BMC Bioinformatics

RESEARCH

Open Access

scHiCSRS: a self-representation smoothing method with Gaussian mixture model for imputing single cell Hi-C data

Qing Xie[1], Wang Meng[2] and Shili Lin[1,3*]

*Correspondence:
shili@stat.osu.edu

[1] Interdisciplinary Ph.D. Program in Biostatistics, The Ohio State University, Columbus, OH 43210, USA
[2] Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH 43205, USA
[3] Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

Abstract

Background: Single cell Hi-C (scHi-C) techniques make it possible to study cell-to-cell variability, but excess of zeros are makes scHi-C matrices extremely sparse and difficult for downstream analyses. The observed zeros are a combination of two events: structural zeros for which two loci never interact due to underlying biological mechanisms, or dropouts (sampling zeros) where two loci interact but not captured due to insufficient sequencing depth. Although data quality improvement approaches have been proposed, little has been done to differentiate these two types of zeros, even though such a distinction can greatly benefit downstream analysis such as clustering.

Results: We propose scHiCSRS, a self-representation smoothing method that improves data quality, and a Gaussian mixture model that identifies structural zeros among observed zeros. scHiCSRS not only takes spatial dependencies of a scHi-C data matrix into account but also borrows information from similar single cells. Through an extensive set of simulation studies, we demonstrate the ability of scHiCSRS for identifying structural zeros with high sensitivity and for accurate imputation of dropout values in sampling zeros. Downstream analyses for three experimental datasets show that data improved from scHiCSRS yield more accurate clustering of cells than simply using observed data or improved data from comparison methods.

Conclusion: In summary, scHiCSRS provides a valuable tool for identifying structural zeros and imputing dropouts. The resulted data are improved for downstream analysis, especially for understanding cell-to-cell variation through subtype clustering.

Keywords: Structural zeros, Dropouts, Sampling zeros, Neighborhoods, Sparsity

Introduction

The spatial organization of chromosomes in a cell nucleus is not random; rather, it is dynamic and closely linked to genome functions and disease mechanisms [4]. Harnessing the power of next-generation sequencing technologies, the Hi-C technology enables a high resolution, genome-wide three-dimensional (3D) view of the chromosomal organization [15], and it has been applied to analyze different types of cells [3, 12, 24]. The original Hi-C technique produces bulk data, averaging chromosome

conformation over millions of cells and resulting in limited information on cell-to-cell variability [6]. Recent single cell Hi-C assays, on the other hand, enable the analysis of whole-genome structures for single cells [18] and has the potential to identify rare cell populations or cell sub-types in a heterogeneous population [22].

Interpreting single cell Hi-C (scHi-C) data is challenging because of data sparsity (observed zeros) and low sequencing depth [19]. Due to the increase of data dimension, the coverage of scHi-C (0.25–1%) is much smaller than that of RNA-seq (5–10%) [38], leading to additional difficulty for analyzing scHi-C data. The observed zeros are a mixture of two types of events: some are structural zeros because the pairs do not interact with each other due to the underlying biological mechanisms, while others are dropouts or called sampling zeros as a result of low sequencing depth. While dropouts happen at random, structural zeros do not. Differentiating between structural zeros and dropouts and imputing the latter can lead to improved downstream analyses such as clustering and 3D structure inference.

The zero-inflated phenomenon is also observed in single cell RNA (scRNA) research. Currently, there is considerable research on imputation for scRNA data, with the concept of structural zero well defined and inferences made to distinguish structural zeros and dropouts [1, 9, 14, 17, 21, 23, 28, 32, 39]. In contrast, the concept and inference of structural zeros and dropouts have not been widely discussed in scHi-C research, although we note that in several papers that aim to assess data reproducibility [27, 31], construct 3D structures [40], cluster single cells [38], or infer pseudotime path [16], imputing values for observed zeros has been treated as an intermediate data enhancing step. In a recent contribution, we explored the potential of using scRNA methods for analyzing scHi-C data and achieved some success [7]. However, the issue of scRNA methods not accounting for spatial correlation - a hallmark of Hi-C data - was also identified.

In the Hi-C literature for quality improvement for bulk or single cells data, kernel smooth, random walk, and convolutional neural network are the main ideas [16, 27, 31, 38, 40]. The 2D mean filter approach (a kernel smoothing method) directly replaces each cell of a 2D contact matrix with the mean count of all contacts in its genomic neighborhood. For example, HiCRep [31] applies such a filter to assess the reproducibility of Hi-C data, while a recent update, ScHiC-Rep [36], applies a uniform kernel to cluster scHi-C data. scHiCluster [38] applies a convolution-based imputation including a mean filter to help cluster cells. Different from a 2D mean filter that takes an average of the genomic neighbors, kernel smooth uses a weighted average of neighboring observed counts. The weight is defined by a kernel, which gives more weight to closer genomic neighbors. For instance, SCL [40] applies a 2D Gaussian function to impute scHi-C contact matrices and further infers the 3D chromosome structures from the enhanced Hi-C data. GenomeDISCO [27], on the other hand, uses a random walk on the contact map to "smooth" the observed counts, and it shows that taking three steps of the random walk would lead to best results in general. scHiCluster [38] also uses the idea of a random walk, but with restarts, to capture the topological structure. Convolutional neural network is also an approach commonly applied to infer a high-resolution Hi-C matrix from a low-resolution one. HiCPlus

[33] and DeepHiC [8] are examples of such supervised learning techniques. Examples of combining linear convolution and random walk also exist [16, 38]

Although taking spatial correlation in a 2D data matrix into consideration, the current methods as discussed above enhance each Hi-C data matrix independently without considering other information, such as data from similar cells. Further, inference on structural zeros and dropouts is rarely discussed, although the identification of such may play an important role in downstream analyses. In an attempt to make fuller usage of available information and to distinguish structural zeros from dropouts, in this paper, we develop scHiCSRS, a self-representation smoothing method. It not only borrows information from 2D neighborhoods but also takes similar single cells into account. Further, as part of the scHiCSRS package, we propose a Gaussian mixture model to separate the zeros into structural zeros and dropouts. Through an extensive set of simulation studies and experimental data analyses, we showed that scHiCSRS can accurately identify structural zeros and impute the dropouts. We also compared scHiCSRS with other methods for data quality improvement and downstream clustering analyses.

Methods

The scHiCSRS workflow

The overall goal of scHiCSRS is to enhance scHi-C data and make inference on structural zeros (Fig. 1). scHiCSRS takes spatial dependencies of scHi-C 2D data structure into consideration while also borrows information from similar single cells. scHiCSRS was motivated by scTSSR [11] that recovers scRNA data using a two-sided sparse self-representation method, but there are two major differences. Firstly, scTSSR uses the expression of all genes in the same cell while scHiCSRS only considers counts in a 2D matrix neighborhood, which helps capture local dependencies [36]. Secondly, scTSSR

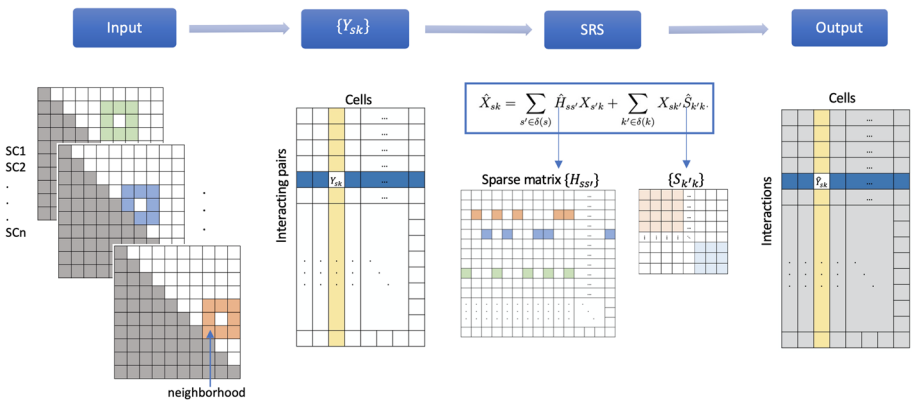

Fig. 1 Schematic of the scHiCSRS algorithm. *Input:* the input includes multiple scHi-C contact matrices, with the colored region of a cell denoting the neighborhood of a position enclosed. *Data matrix* $\{Y_{sk}\}$: the single cells are organized into a big matrix, with each row representing a pair of interacting loci and each column being the upper triangular part of a single cell contact matrix. *SRS:* A self-representation model is used to enhance the entries in the matrix X (normalized from the observed matrix Y); since SRS only borrows information from 2D neighborhoods, the coefficient matrix H is sparse with its values in most positions (not in the neighborhood of a position) set to 0; if the input single cells are composed of more than one type, the matrix S is also sparse, with only non-zero blocks along the diagonal because we only consider the influence from similar single cells organized as groups. *Output:* the output is the enhanced matrix $\{\hat{Y}_{sk}\}$, based on which we can perform additional analyses

has an interaction term that involves elements in the same row and column; however, scHiCSRS does not include such a term because other positions in other single cells should have no direct influence on the position to be imputed. Based on the quality-improved data, we further apply a Gaussian mixture model to identify structural zeros, as details unfold in the following.

Self-representation smoothing model

Suppose we have contact matrices for K single cells. Let Y_{ijk} represents the observed interaction frequency between loci i and j $(i \le j)$ for single cell k $(k = 1, \cdots, K)$, where a locus is a gnomic segment and $\{Y_{ijk}\}_{n\times n}$ is a symmetric 2D matrix of dimension $n \times n$ for each single cell $k, 1 \le k \le K$, where K is the number of single cells and n is the number of genomic loci considered. We combine the 2D contact matrices of all single cells into a big matrix $\{Y_{sk}\}$ $(s = 1, \cdots, N = n(n+1)/2, k = 1, \cdots, K)$ of dimension $N \times K$ with each column being the upper triangular (including the diagonal) of a single cell 2D matrix. We first normalize each cell so that all cells have the same sequencing depth (the median---med---across all cells), then we log-transform the normalized matrix as follows:

$$X_{sk} = \ln\left[\frac{Y_{sk}}{c_k} + 1\right], s = 1, \cdots, N, k = 1, \cdots, K,$$

where $c_k = \sum_s Y_{sk}/\text{med}\{\sum_s Y_{sk}, k = 1, \cdots, K\}$ is the depth-adjusted normalization factor for cell k, and a pseudo count of 1 is added due to the existance of observed zeros.

For each X_{sk}, there are two types of information that we use for the smoothing process: the neighborhood $\delta(s)$ and the collection of similar cells $\delta(k)$ at the same position; that is, $\delta(s)$ contains the 2D neighbors of position s (but not s itself) while $\delta(k)$ contains all the cells that are similar to k (but not k itself). To smooth the contact matrix, we assume that the contact count of each pair is a linear combination of these two types of information. Therefore, we propose the following self-representation smoothing model for obtaining a "smoothed" scHi-C matrix:

$$X_{sk} = \sum_{s' \in \delta(s)} H_{ss'} X_{s'k} + \sum_{k' \in \delta(k)} X_{sk'} S_{k'k} + \epsilon_{sk},\ s = 1, \cdots, N, k = 1, \cdots, K, \tag{1}$$

where the $\{H_{ss'}\}_{N\times N}$ and the $\{S_{k'k}\}_{K\times K}$ matrices are described as follows, and ϵ_{sk} is the error term.

For convenience, the neighborhood $\delta(s)$ is taken to be a regular one, as shown in Fig. 1, although the size and shape may be modified as appropriate (see Discussion). For all the data analyses carried out in this paper, we use a regular neighborhood with 24 neighbors. This is a 5×5 squared neighborhood, with the position of interest occupying the center of the lattice. The $N \times N$ matrix $H = \{H_{ss'}\}_{N\times N}$ describes the influence of neighbor s' on position s so that only positions within the neighborhood have a positive coefficient and the others are set to 0, leading to a sparse matrix (Fig. 1). Since we work with a vectorized upper-triangular contact matrix, the neighborhood information is transformed into the vector, and subsequently the H matrix, according to how the matrix is vectorized (e.g. by

row, by column, or by diagonal band). However, we note that results are not influenced by how the contact matrix is vectorized, since such information is adjusted accordingly in H.

The $K \times K$ matrix $\{S_{k'k}\}_{K\times K}$ describes the influence of cell k' on cell k and is set in such a way that only similar cells $k' \in \delta(k)$ have a positive influence, the rest is set to 0. Thus, if the input single cells consist of different types, then the matrix $S_{k'k}$ would have non-zero blocks along the diagonal with each block being the coefficients for single cells of the same type (Fig. 1). At first glance, the first term in the right-hand side of Eq. (1) is similar to kernel smoothing methods, where values from a well-defined neighborhood of a position in the 2D matrix are borrowed for imputation. However, there is a key difference between the two: while the weights in kernel smoothing are pre-determined by the kernel used (e.g. uniform kernel or Gaussian kernel), they are estimated internally (without specifying a kernel) based on the observed 2D matrices. This level of flexibility allows scHiCSRS to find the best estimates of the weight matrix H using the data at hand. The additional feature that differentiates scHiCSRS from kernel smoothing method is the inclusion of the second term in (1), where information from cells of the same type is also utilized. In other words, cells of the same type, if such information is available, will be used in our grouping for setting the sparsity in the S matrix. Therefore, although all cells can be analyzed together, cells from different types are not mixed. This is reflected in our sparse S matrix, where it is flexibly modeled so that only cells of the same type contribute to the imputed values, and the weights are flexibly estimated rather than being fixed.

We estimate the coefficient matrices $H = \{H_{ss'}\}$ and $S = \{S_{k'k}\}$ in the self-representation smoothing model through a penalized least squared method [11]. We define the following objective function:

$$f(H,S) = ||X - (HX + XS)||_F^2 + \lambda||S||_1,$$

where $||\cdot||_F$ and $||\cdot||_1$ are the Frobenius and the l_1 norm, respectively, and λ is a non-negative tuning (penalty) parameter. Thus, this may be interpreted as analogous to a Lasso-type objective function. According to Gordon's Theorem [29], a proper Lasso penalty parameter λ is in the order of the standard deviation of the noises [35]. For simplicity and following the literature [11], we fix an estimate for λ before estimating the coefficient matrices. Specifically, we used $X - mean(X)$ to estimate the noise matrix and set the tuning parameter as $\lambda = sd(X - mean(X)) = sd(X)$. Technical details of the optimization procedure can be found in the Supplementary Material. Once we obtain their estimates, denote as $\{\hat{H}_{ss'}\}$ and $\{\hat{S}_{k'k}\}$, respectively, the imputed value is calculated as

$$\hat{X}_{sk} = \sum_{s'\in\delta(s)} \hat{H}_{ss'}X_{s'k} + \sum_{k'\in\delta(k)} X_{sk'}\hat{S}_{k'k}.$$

Although the imputed value $\hat{X}_{sk}$ borrows information from the contacts in neighboring positions in the same cell and other cells at the same position, it does not take the observed value X_{sk} itself into consideration directly. Therefore, we couple the above procedure with the idea of a Bayesian model for scRNA data [10]. Specifically, we model the observed count Y_{sk} (without normalization or log-transform) as follows: $Y_{sk} \sim Poisson(c_k\lambda_{sk})$ and $\lambda_{sk} \sim Gamma(\alpha_{sk}, \beta_{sk})$, where λ_{sk} represents the normalized (med) true interaction intensity and c_k is the normalization factor as defined above.

The marginal distribution of Y_{sk} is then a negative binomial, allowing for over-dispersion. The prior mean (for the Gamma distribution at the normalized scale) is set to be $\hat{\mu}_{sk} = exp(\hat{X}_{sk})$. The prior variance is estimated by maximizing the likelihood assuming independence and common variance across all cells and plugging in the estimated prior mean. Reparameterization leads to the estimated shape and rate parameters, $\hat{\alpha}_{sk}$ and $\hat{\beta}_{sk}$. The posterior distribution is then $\lambda_{sk}|Y_{sk}, \hat{\alpha}_{sk}, \hat{\beta}_{sk} \sim Gamma(Y_{sk} + \hat{\alpha}_{sk}, c_k + \hat{\beta}_{sk})$. We use the posterior mean to estimate λ_{sk} as follows:

$$\hat{\lambda}_{sk} = \frac{Y_{sk} + \hat{\alpha}_{sk}}{c_k + \hat{\beta}_{sk}} = \frac{c_k}{c_k + \hat{\beta}_{sk}} \frac{Y_{sk}}{c_k} + \frac{\hat{\beta}_{sk}}{c_k + \hat{\beta}_{sk}} \hat{\mu}_{sk},$$

which is a weighted average of the normalized observed contact counts and the prior mean estimated from SRS. The final imputed value for Y_{sk}, in the original scale, is $\hat{Y}_{sk} = c_k \hat{\lambda}_{sk}$

Gaussian mixture model

Since the self-representation smoothing model does not have an internal mechanism for separating structural zeros from dropouts, we further propose a Gaussian mixture model on the imputed data $\hat{Y}_{sk}$ to address this issue. We start by normalizing the imputed matrix to the median library size and taking the $\log_{10}$ transformation with a pseudo count 1, the same as described in section 2.2, albeit it is now with the imputed, not the raw counts:

$$Z_{sk} = \log_{10} \left[\frac{\hat{Y}_{sk}}{\sum_s \hat{Y}_{sk}} \times \text{med} \left\{ \sum_s \hat{Y}_{sk}, k = 1, \cdots, K \right\} + 1 \right].$$

We note that the Z_{sk}'s are not counts nor zeros.

Without loss of generality, we assume all the cells are of the same type so that we can use the notation already defined above. If there are multiple known types, then the Gaussian mixture model will be applied to each separately. For a pair of loci (i.e. a position in the 2D Hi-C data matrix) that has zero interaction counts in all the single cells, they are automatically labeled as structural zeros without being subjected to the mixture analysis. For the remaining pairs with zeros in some cells and nonzeros in other cells, collectively denoted as $\mathcal{S}$, we assume that

$$Z_{sk} \sim \eta^1 N(\mu^1, \sigma^1) + \cdots + \eta^G N(\mu^G, \sigma^G), s \in \mathcal{S}, k = 1, \cdots, K,$$

where $\sum_{g=1}^{G} \eta^g = 1$ and $\mu^1 < \mu^2 < \cdots < \mu^G$. That is, the imputed values at the positions with observed zero in some cells follow a G-component Normal mixture distribution. For a position in a cell that has high imputed interaction frequencies, captured by a component with a higher mean, an observed zero is more likely a dropout; whereas if the imputed interaction frequency is low, captured by a component with a lower mean, then an observed zero may be a true structural zero. The parameters are estimated using the Expectation-Maximization (EM) algorithm for a given G, and the best $G, \hat{G}$, is selected based on BIC [2]. We then calculated P_{sk}^{SZ}, the probability of being structural zero for each position $s \in \mathcal{S}$ in each single cell k as follows:

$$P_{sk}^{SZ} = \frac{\sum_{g:g\in R} \hat{\eta}^g f_g(Z_{sk};\hat{\mu}^g,\hat{\sigma}^g)}{\hat{\eta}^1 f_g(Z_{sk};\hat{\mu}^1,\hat{\sigma}^1)+\cdots+\hat{\eta}^{\hat{G}} f_{\hat{G}}(Z_{sk};\hat{\mu}^{\hat{G}},\hat{\sigma}^{\hat{G}})},$$

where the f's are the normal density functions, and R is the Gaussian components designated as the structural zero component(s) based on the following rule. If $\hat{G}=2$, the first component is chosen to capture structural zeros. If $\hat{G}\geq 3$, denote the distances between adjacent means to be $d_{j(j+1)}=\hat{\mu}_{j+1}-\hat{\mu}_j, j=1,2,\cdots,\hat{G}-1$. If $\xi d_{12}\leq d_{23}$ for a large multiple ξ (say, $\xi=10$), meaning that the first two components are close to each other but are far away from the third component, we choose the first and second as structural zero components; otherwise, only the first component is treated as capturing structural zeros. If both of the first and second components are already chosen as capturing structural zeros, we continue the process using the same criterion to ascertain whether additional successive components, up to $\hat{G}-1$, should be chosen. We note that, although ξ is set arbitrarily to be a large number equal to 10, it in fact has little impact on the outcome, since the distance between the first and second ordered components are typically much larger than the subsequent distances. Finally, an observed zero is classified as a structural zero if $P_{sk}^{SZ}\geq 0.5$, although other threshold values may also be considered.

Performance evaluation criteria

For simulation studies, since the "ground truth" is known, i.e., which positions in the 2D matrix are structural zeros and which observed zeros are dropouts, we consider several criteria that make use of such ground truth to evaluate the performance of scHiCSRS and compare it with the other data quality improvement methods. We note, though, that none of the methods, including scHiCSRS and its comparison methods, use the "ground truth" information in the modeling and structural zeros/dropouts inference. First, we evaluate the ability of scHiCSRS to identify structural zeros among the observed zeros, and to compare its performance with methods in the literature. Specifically, for the comparison methods, since they do not have an internal mechanism for identifying structural zeros, we label an observed zero as a structural zero if the imputed value is less than 0.5, following suggestions in the literature [7]. To measure the ability of a method (scHiCSRS or a comparison method) for separating structural zeros from sampling zeros, we call the proportion of true structural zeros identified as the *power* or *sensitivity*, defined as the proportion of underlying structural zeros correctly identified. Similarly, we call the proportion of true dropouts, defined as the proportion of underlying sampling zeros correctly identified, as the *specificity* to measure the ability of a method for correctly identifying dropouts. Since the identification of structural zeros and dropouts depends on the decision rules (a threshold on the probability for the Gaussian mixture model or a threshold on the imputed value for the comparison methods), we also explore a range of thresholds, with the result measured as the area under the curve (AUC), the receiver operating characteristic (ROC) curve, for a more thorough comparison of methods. We use the absolute errors between the imputed and the expected values to further assess the imputation accuracy of scHiCSRS and the comparison methods. Moreover, we use the correlations between the imputed and the expected to measure the aggregate performance of a method to assess the imputation accuracy. In addition to those criteria,

scatterplots and 2D heatmaps are used to visualize the imputation performance. For experimental data analysis, since expected and SZ designations of observed zeros are unknown, our analysis of the scatterplot and correlation is restricted to non-zero observations. We further use the adjusted rand index (ARI) to evaluate the clustering results.

Results

Simulation study

Data generation

To mimic experimental data, we use three 3D structures on a segments of chromosome 1 (the first 61 mega bases loci) recapitulated using SIMBA3D [25] from three K562 single cell Hi-C 2D matrices [5]. For each structure (single cell), based on the estimated 3D coordinates $(x_i, y_i, z_i)\,(1 \le i \le 61)$, we firstly generate the interaction intensity matrix $\lambda = \{\lambda_{ij}\}$ with the following model:

$$\log(\lambda_{ij}) = \alpha_0 + \alpha_1 \log d_{ij} + \beta_l \log(x_{l,i}x_{l,j}) + \beta_g \log(x_{g,i}x_{g,j}) + \beta_m log(x_{m,i}x_{m,j}), 1 \le i \le j \le 61,$$

where $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$ is the distance between loci i and j; $x_{l,i} \sim \text{Unif}(0.2, 0.3)$, $x_{g,i} \sim \text{Unif}(0.4, 0.5)$, and $x_{m,i} \sim \text{Unif}(0.9, 1)$ mimic covariates such as fragment length, GC content, and mappability score [20], and β_l, β_g, and β_m are the corresponding coefficients of the covariate terms; α_1 is set to -1 following the typical biophysical model; and α_0 is a scale parameter that we used to control sequencing depths.

These three structures are designated as three "types" (I, II, and III) of single cells. For each type, we simulate n single cells, with varying numbers of n's as described below. To simulate sparse 2D matrices with both structural zeros and dropouts, we define a threshold b as the lower 10% quantile of the λ_{ij}'s. For those $\lambda_{ij} < b$, we randomly select half of them to be structural zeros candidates; among them, 80% are randomly selected to be structural zeros across all n single cells. For a particular single cell, we randomly select half of the remaining 20% candidates to be structural zeros. For those candidates that are selected as structural zeros, their new λ_{ij} are set to be zero while for those that are not selected to be structural zeros, the λ_{ij} values are left unchanged in the original λ matrix. This procedure makes each single cell has its own specific λ^* matrix (containing "expected" values). Based on the λ^* matrix, we generate the contact counts using a Poisson distribution with the intensity parameter being the corresponding λ^*_{ij} for a particular single cell. This step also produces dropouts that are observed zeros but their underlying true values are nonzero. Using three sets of parameters (Table S1), we simulated single cells for type I, II, and III with three sequencing depths (7k, 4k, and 2k) and three sample sizes of cells (10, 50, 100). This data generation process and the parameters chosen follow those in the literature [7].

Simulation study results

In addition to scHiCSRS, we also considered four other methods that have been used as an intermediate step to enhance Hi-C data for comparison. These four methods represent major categories of methods discussed above. They are mean filter (MF) as in HiCRep [31], which replaces each contact with the average count of its neighborhood region; Gaussian kernel smooth (GK) as in SCL [40], which uses a weighted average of neighboring observed data and the weights are determined by a Gaussian kernel;

random walk (RW) as in GenomeDISCO [27], which takes a 3-step random walk; and convolution and random walk (CR) as in scHiCluster and scHiCPTR [16, 38], which combines linear convolution with random walk. Since these comparison methods do not have a built-in mechanism for identifying SZs, we borrow a criterion from the existing literature on scRNA-seq [17, 28] and recently in scHi-C [7] that labels a position as structural zero if the imputed value is less than 0.5.

We focus on two major tasks for evaluating scHiCSRS and comparing the performance with those of the four selected methods. Our first and foremost objective is the ability for the methods to correctly separate SZs from the DOs, where a good method is one that has a high sensitivity (power) of identifying SZs without including, incorrectly, a significant portion of DOs, that is, with high specificity. The second focus is to evaluate the ability of the methods to accurately impute the DOs, evaluated by the criteria discussed above, including absolute difference and correlation between the imputed and expected values, aided further by visual inspection using scatterplots.

Sensitivity for detecting SZs. For correct identification of SZs (i.e. sensitivity), we see that in all settings considered regardless of the sequencing depth (7k, 4k, or 2k), sample size (number of cells: 10, 50, or 100 cells), and cell type (I, II, or III), scHiCSRS has a power of near 0.9 or higher (Fig. 2a, and Table S2).

In contrast, the performance of the four comparison methods fluctuates greatly with sequencing depth: it may be as high as 0.84 when the sequencingx depth is 2k, but may be down to zero when the sequencing depth is 7k.

We also used ROC curves to explore the interplay between correct identification of structural zeros and dropouts for a fair comparison of all methods, especially since the criterion for declaring SZs may perhaps be viewed as arbitrary despite using guidelines from the literature. The ROC curves of scHiCSRS go up to 1 quickly (Fig. 2b) and the AUC are at least 0.85 for all combinations of sequencing depth, cell types, and number of cells (Table S3). The corresponding ROC curves for the four comparison methods, on the other hand, lie much below their scHICSRS counterpart, with the AUC as low as 0.52. These results indicate that scHiCSRS has both higher sensitivity and specificity than the four comparison methods.

Since structural zeros are critical for downstream analysis such as 3D structure construction [30, 34], we are also interested in evaluating the performance of the methods when the proportion of correctly identified true structural zeros, the power, is kept at a high level. As such, we compare the performance of the four methods when the power is fixed at 0.95. For every combination of cell type, sample size, and sequencing depth, scHiCSRS maintains a much higher specificity for correctly identifying true dropouts (Fig. 2c and Table S4). Although the performance is not even for all combination of scenarios considered, as the specificity may dip down to 0.72 for type II cells at sample size 10 with 7k sequencing depth, scHiCSRS still outperforms the comparison methods by a big margin, as the best performer among the four comparison methods only obtained a specificity of 0.26. In general, one can see that the performance of the four comparison methods is not sensitive to the number of cells, but is sensitive to the sequencing depth. In particular, for types II and III, the proportions are much smaller when the sequencing depth is 7k.

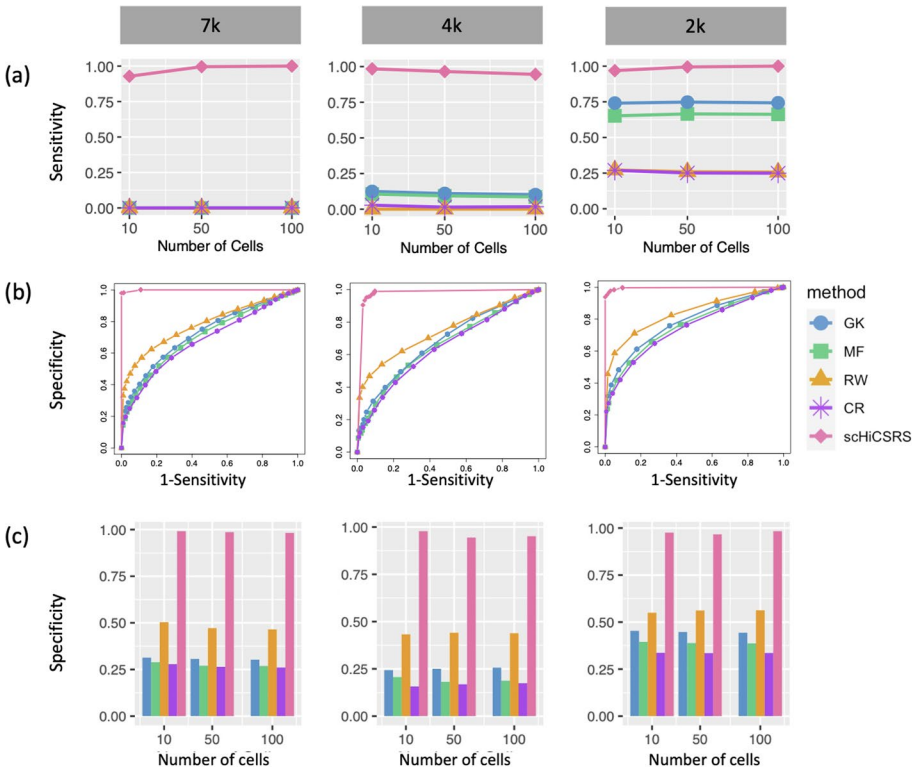

Fig. 2 Performance evaluation (average over 100 replicates) of scHiCSRS and comparison methods for type I cells with three sequencing depths: 7k (1st column), 4k (2nd column), and 2k (3rd column). **a** Sensitivity for detecting structural zeros; **b** ROC curves constructed with a range of thresholds for the setting with 100 cells; **c** Proportion of true dropouts correctly detected (specificity) when the detection rate for the proportion of true structural zeros (sensitivity) is set to be 0.95

Imputation accuracy. For assessing imputation accuracy, we consider the correlation and absolute error between the imputed values and the expected values underlying our simulation. We can see that scHiCSRS has the highest correlations compared to the other methods in each of the scenarios studied (first row of Fig. 3a and Table S5). Even though the correlations are overall the lowest for type III cells, the difference still range from 0.13 to 0.23 between scHiCSRS and the best performer among the four comparison methods. Evaluation based on the absolute error shows that it is the smallest for scHiCSRS across cell types, sample size, and sequencing depth (second row of Fig. 3a and Table S6), consistent with the correlation results.

We also visualized the imputation accuracy through the scatterplot between the imputed versus expected values (Figs. 3b and S1-S3). scHiCSRS shows tight clustering of points around the diagonal line, indicating its imputation accuracy. On the contrary, the imputed values of the four comparison methods deviate from what are expected greatly. Their scatterplots end up with funnel shapes due to more variability on larger counts. Further, one can see that, due to a large portion of SZs incorrectly classified as DOs (Fig. 2 and Table S4), the scatterplots for the four comparison methods show a rather obvious red line segment at the zero expected value. In other word, the unidentified true SZs were imputed with values covering a significant range. This phenomenon is not seen in the scHiCSRS scatterplot. The rest of the red dots with

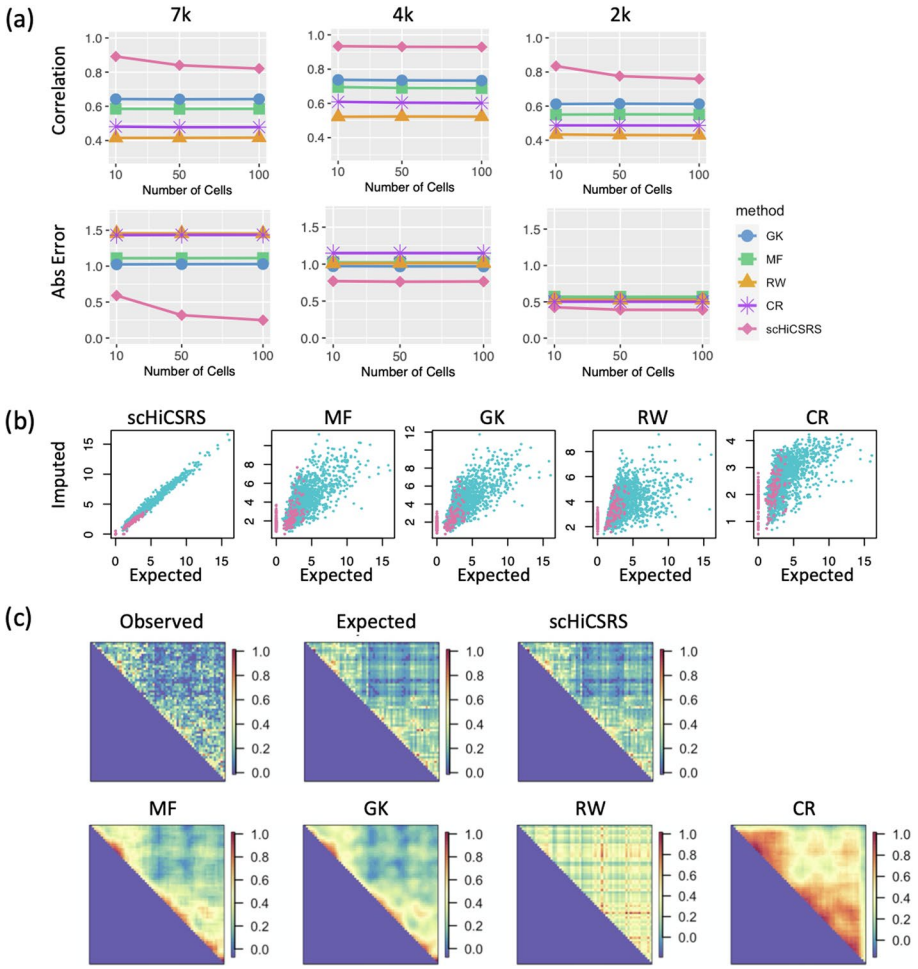

Fig. 3 Comparison of imputation accuracy of scHiCSRS and comparison methods. **a** Correlation between imputed values and expected (row 1) and absolute difference between imputed and expected (row 2) for type I cells over three sequencing depths and three sample sizes (number of cells), averaging over 100 replicates: 7k (1st column), 4k (2nd column), and 2k (3rd column); **b** Scatterplots of expected versus imputed for the first type I single cell with 7k sequencing depth: the red dots represent observed zeros, which contain both true SZs (expected = 0) and DOs; **c** Heatmaps of the first type I single cell with 7k sequencing depth showing the data (Observed) and true (Expected) 2D matrix images as well as the results from scHiCSRS and the comparison methods (MF, GK, RW, and CR)

non-zero expected values, representing true DOs, also show large scattering patterns for the four comparison methods, indicating inaccuracy in the imputed values, compared to the tight clustering around the diagonal line for scHiCSRS.

As a further illustration of the ability of scHiCSRS for recovering the underlying structures of chromatin interactions, we visualize the 2D heatmaps of the contact matrices (Fig. 3c). In the first two heatmaps, we can see that the clean underlying (expected) structure is blurred by the existence of sampling zeros that are a mixture of SZs and DOs in the observed image. After processed by scHiCSRS, the image is almost entirely restored, clearly showing two TADs with well-defined boundaries and correct relative magnitudes of contacts. In contrast, the images recovered by MF, GK, and CR are blurry. Although the boundaries of the two TADs are clearly demarcated,

the relative magnitudes within the domains are rather different from the expected, corroborating our finings when assessing the accuracies of the imputed values. Finally, the heatmap of RW was unable to recover the domain structure, not surprisingly due to the underlying Markovian property of converging to identical rows, even just with three steps.

Experimental data analyses and results

We consider three experimental scHi-C datasets to demonstrate the improvement of downstream analysis after data improvement with scHiCSRS and compare with the results using data improved by the four comparison methods: MF, GK, RW, and CR. For all three datasets, the bin size was set to be 1 MB due to the limited sequencing depth in the data. We first explored whether the imputed data from scHiCSRS and the four comparison methods can improve downstream clustering using the K-means algorithm and assessed the results based on the adjusted rand index (ARI). As mentioned earlier, for experimental data, the underlying SZs and the expected values for DOs are unknown; therefore, our assessment of the quality of imputed data relies on those that have non-zero observed values, essentially using our scatterplots and correlation analysis tools for observed non-zeros. Nevertheless, the imputed values for the non-zero observations should not deviate wildly from their observed counterparts. Thus, the scatterplots and correlations between the imputed and non-zero observed values were used further to indirectly assess whether the imputed values were sensible for the observed zeros.

GSE117874 - GM vs. PBMC

Our first experimental data analysis was on a dataset (GSE117874) that consists of 14 lymphoblastoid cells (GM) and 18 peripheral blood mononuclear cells (PBMC) [26], download from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117874. We used a sub-2D matrix of dimension 30×30 on chromosome 1 for demonstrating the results before and after imputation (Table 1a). We restricted ourself to only a subset of the available data so that we can delineate the performances of the various methods. The analysis based on the observed data led to a total of eight misclassifications: one GM was in cluster C2 with the majority of the PBMC cells, while seven of the PBMCs were in cluster C1 with the GM cells. Note that, in this dataset, since the two cells types are clearly different and the identity of each of the 32 single cells are known, we can evaluate whether clustering results from the observed data (with a total of eight misclassifications) can be improved after imputation using scHiCSRS and the comparison methods.

The number of zero positions identified to be structural zeros were 154 among the 14 GM cells and 234 among the 18 PBMC cells. Together with the imputed dropout values led to the scHiCSRS-enhanced data, which corrected two of the misclassified PBMC cells (Table 1a), leading to an increase in ARI (Fig. 4a Column 1). On the other hand, MF and GK did not result in any improvement - still having eight misclassifications; RW led to misclassifications of two more GM and one more PBMC cells; and CR corrected one of the misclassified PBMC cells but misclassified an additional four GM cells. Hence, the ARI remains unchanged for MF and GK, while RW and CR lead to much smaller ARI's. We then computed the correlation between the imputed values and the observed non-zeros for each single cell, and one observes

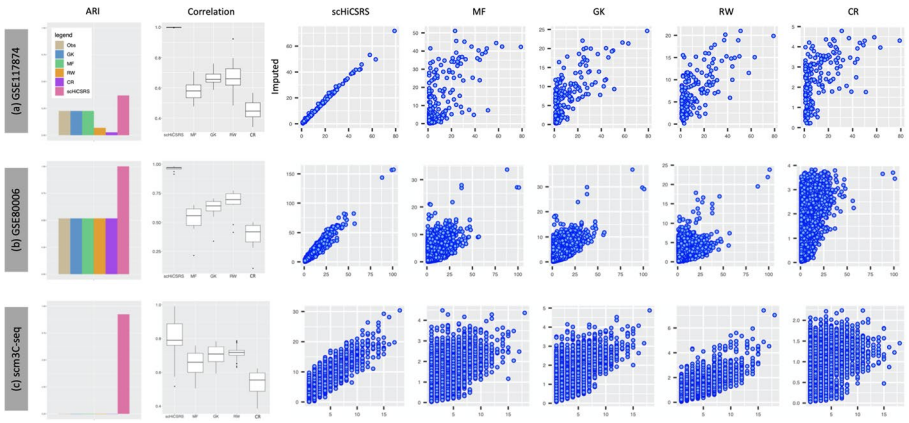

Fig. 4 Comparison of five methods on three experimental datasets: **a** GSE117874; **b** GSE80006; **c** scm3C-seq. Column 1: adjusted rand index (ARI) for assessing clustering; Column 2: boxplots of correlations between the imputed and nonzero observed values with the number of cells in each dataset; Columns 3–7: scatterplots of imputed versus nonzero observed values for a randomly selected cell with each of the five methods studied

Table 1 K-means (KM) Clustering results for three single-cell Hi-C datasets with the observed and improved data from five different methods

CellType	KM clusters	Observed	scHiCSRS	MF	GK	RW	CR
(a) GSE117874 - GM vs. PBMC							
GM	C1	13	13	13	13	11	9
	C2	1	1	1	1	3	5
PBMC	C1	7	5	7	7	8	6
	C2	11	13	11	11	10	12
(b) GSE80006 - K562A vs. K562B							
K562A	C1	1	2	1	1	1	1
	C2	1	0	1	1	1	1
K562B	C1	0	0	0	0	0	0
	C2	8	8	8	8	8	8
(c) scm3C-seq - L4 cv. L5							
L4	C1	76	131	77	77	76	125
	C2	55	0	54	54	55	6
L5	C1	105	6	105	104	105	171
	C2	75	174	75	76	75	9

high correlations for all cells with the scHiCSRS-imputed data, exhibited as a tight box around 1 (Fig. 4a Column 2). In comparison, the correlations are much smaller, with much more cell-to-cell variability, for the data imputed from the other four methods. To more fully understand the results, we plotted the imputed versus the observed for non-zero observed values. One can see that, for scHiCSRS, the points are tightly arranged on a straight line (Fig. 4a Column 3), explaining the extremely high correlation. In contrast, the point clouds scatter loosely with an obvious funnel shape for each of the four comparison methods (Fig. 4a Columns 4-7). Not revealed by simply computing the correlation, the plots show that the scHiCSRS-enhanced matrix has an appropriate range: the range of the imputed values is comparable to that of the observed non-zero values, indicating that the imputed values are likely to

be reasonable for the observed zeros. However, MG, GK, RW, and CR led to substantially narrower range for the imputed data, especially CR, likely due to the non-zero observations being "neutralized" by their sparse neighborhoods when repeatedly applying random walk until convergence. This phenomenon was also observed for the simulated data, although the results were not as extreme.

GSE80006 - K562A vs. K562B

Our second experimental data analysis was on a dataset (GSE80006) that consists of scHi-C data of 19 K562A and 15 K562B cells [5], downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80006. Our analysis considered all the intra-chromosomal data of 10 single cells, two K562A and eight K562B, that had a sequencing depth of at least 5K. For the rest of the 24 single cells, their sequencing depths were all around just 1K, too little to provide sufficient information for data enhancement, even at the 1 MB resolution. For the 10 cells with reasonable sequencing depth, all but one were correctly classified using the observed data before any data enhancement. The only misclassification was a K562A cell, which was grouped with the eight K562B cells in cluster C2 (Table 1b).

For data enhanced with scHiCSRS, the misclassified K562A cell was now correctly grouped with the other K562A cell in cluster C1. The eight K562B cells remained clustered together in C2, leading to a perfect ARI of 1.0. The ability of scHiCSRS for correctly classifying the K562A cell could be due to the discrepancies in the percentages of structural zeros identified: for the K562A cells, the average percentage of structural zeros was 91% (SD = 6.2%); for K562B, the average was much lower at 65% (SD = 2.8%). On the other hand, MF, GK, RW, and CR all failed to correct the misclassified cell, and thus with the ARI remaining at 0.5 (Table 1b and Fig. 4b Column 1). Not surprisingly, the correlation between the imputed and the observed non-zero values are extremely high for all single cells with scHiCSRS-improved data, whereas the counterparts for MF, GK, RW, and CR are all considerably smaller and with larger variability across cells (Fig. 4b Column 2). The high correlations for scHiCSRS are once again reflected in the scatterplot of the observed versus the imputed, as the points are clustered tightly on a diagonal line, although the range for the imputed data are a bit larger than that for the observed (Fig. 4b Column 3). In contrast, there is a great deal of scattering for the data imputed with MF, GK, RW, and CR, and the imputed data are all in much smaller ranges - less than 1/2 - of that of the observed (Fig. 4b Columns 4-7).

scm3C-seq - L4 vs. L5

Our third experimental data analysis was on a dataset (scm3C-seq) consisting of scHi-C data of human brain prefrontal cortex cells, downloaded from https://github.com/dixonla b/scm3C-seq. Our analysis focused on two neuronal subtypes, L4 and L5, which are known to locate on two different cortical layers. However, these two subtypes were all mixed together using the observed scHi-C data in a previous study [13], although they were shown to be two separate subtypes using DNA methylation data. Thus, this represents a much more challenging problem than the first two datasets. In our analysis, we considered intra-chromosomal data of 131 L4 cells and 180 L5 cells. Our own clustering analysis with the observed data confirms the finding of the previous study that

the observed scHi-C data were incapable of separating these two subtypes (Table 1c). Specifically, 76 of the L4 and 105 of the L5 cells were clustered into C1, whereas 55 of the L4 and 75 of the L5 were clustered into C2, a highly mixed result leading to an ARI of near zero. We further explored whether the results would improve with a larger number of clusters. Our additional analysis showed that increasing the number of clusters would not necessarily lead to an improvement of the result since the "elbow" of the clustering objective measure, the within-cluster sum of squares, occurred when the number of cluster was 2 (Figure S4).

Using data enhanced with scHiCSRS, we can see that the clustering result is greatly improved: all the 131 L4 cells were clustered together into C1, and all but six of the 180 L5 cells were clustered together into C2. This led to a much improved ARI of over 0.9 (Table 1c and Fig. 4c Column 1). There are good correlations between the observed and the imputed data for most of the single cells, although the results are not as good as those for the first two data sets, reflecting the difficulty of analyzing this third dataset, especially for several of the single cells with a low correlation. The greater difficulty of handling this dataset is also reflected in the range of the imputed data, which greatly exceeded the observed one. Further, although the point cloud still scatters around a straight line, it is not as tight as with the first two datasets. Using the imputed data from MF, GK, and RW, the clustering results are all practically identical to that using the observed data (Table 1c). For CR, the result in fact becomes much worse. Unsurprisingly, the corresponding ARIs are all near zero; the correlations between the imputed and observed values are generally lower compared to those from scHiCSRS; and there is much more scattering of the imputed versus observed point clouds, especially for those imputed with MF, GK, or CR.

Discussion

This paper proposes a self-representation smoothing method, coupled with mixture modeling, for scHi-C data quality improvement and identification of structural zeros. From both the simulation and the experimental data studies, we can see that scHiCSRS outperforms existing methods for the accuracy of imputing the contact counts of dropouts based on multiple criteria. We can also see that the Gaussian mixture model has the ability to identify components that represent structural zeros in the mixture, which is much better than the comparison methods using thresholding as suggested in the literature. In particular, the results, based on the simulation studies carried out, suggest that structural zeros identification in scHiCSRS is not sensitive to sequencing depth given its adaptive nature, and is also quite stable for the varying number of single cells. We note, though, that incorrect identification of dropouts may lead to worsening performance for some combinations of the cell type, sequencing depth, and the number of cells. In contrast, the comparison methods are rather sensitive to sequencing depth in calling structural zeros using a fixed threshold. Although these comparison methods do not appear to be sensitive to the number of cells since each cell is analyzed separately, the influence of the cell type is much greater. Finally, note that scHiCSRS borrows information from neighbors without pre-specifying the weights but rather estimating them adaptively, which contributes to its better performance in imputation accuracy. Further, the results

do not appear to be sensitive to the shape or size of the neighborhood specification as long as there is a reasonable number of neighbors (Supplementary material).

Through the analyses of three experimental datasets, we further demonstrated that the improved data from scHiCSRS can positively and greatly impact downstream analyses. Using clustering as an example type of downstream analyses, the advantage of scHiCSRS-improved data over the comparison methods is clear. Specifically, from the examples of clustering GM and PBMC cells, K562 cells, and the L4 and L5 prefrontal cortex cells, we have seen that data improved with scHiCSRS led to more accurate clustering, judging from known cell types. For an experimental data analysis, although it is difficult to judge from an absolute sense whether structural zeros were better identified using scHiCSRS, we anticipated this being the case from our extensive simulation results where the ground truth is known. Further, based on the observed non-zero values from the experimental data, we can see that scHiCSRS outperformed MF, GK, RW, and CR in terms of preserving the range of observed values, and thus showing greater potential and credibility for the imputed observed zeros that are in fact dropouts.

Since A/B compartmental information as well as topologically associated domains (TADs) calling information are obtainable for the GM cell line, we sought to understand structural zeros identified by scHiCSRS in terms of their relationships with these two pieces of information. For the GM single cells analyzed, we identified a total of 154 structural zeros, but 112 of them were in a genomic region at the beginning of chromosome 1 that was not assigned to a compartment. Of the remaining 42 structural zeros, the loci involved are all in the inactive B compartment. This result appears to be biological sensible since loci in the B compartment are not expected to be active. In terms of TADs, 126 of the 154 structural zeros are between loci in two different TADs, while the remaining 28 are between two loci with one in a TAD and the other not in a TAD (either in a gap or a boundary region). This result is also biologically interpretable, as we would not expect structural zeros to be between two loci within a TAD.

The much better performance of scHiCSRS is inherently linked to its ability of correctly identifying structural zeros, which plays an important role in clustering. For calling structural zeros using the comparison methods, we rely on a threshold, and our results, based on the ROC curves from the simulation, indicate that the performance will likely be unsatisfactory regardless of the threshold value chosen. This conclusion is not only supported by numerical criteria, but also by examining how close the imputed values are to the corresponding observed values and how well the underlying structures (visualized by heatmaps) can be recovered. One could potentially further utilize the structural zero information for downstream analyses. Suppose there are common structural zero patterns among cells within each unknown subtype yet different patterns across different subtypes, then this information may be leveraged to design novel clustering algorithms that have better discriminating power for identifying subtypes with distinct structural patterns.

The advantages of scHiCSRS not withstanding, there are several issues that deserve further investigation and improvement. First, although it is clearly seen that scHiCSRS outperforms the comparison methods for preserving the range of observed non-zeros (Figure 4 Columns 3-7), the performance on the three experimental data sets is uneven. For GSC117874, the range is completely preserved, with the imputed values tracking the

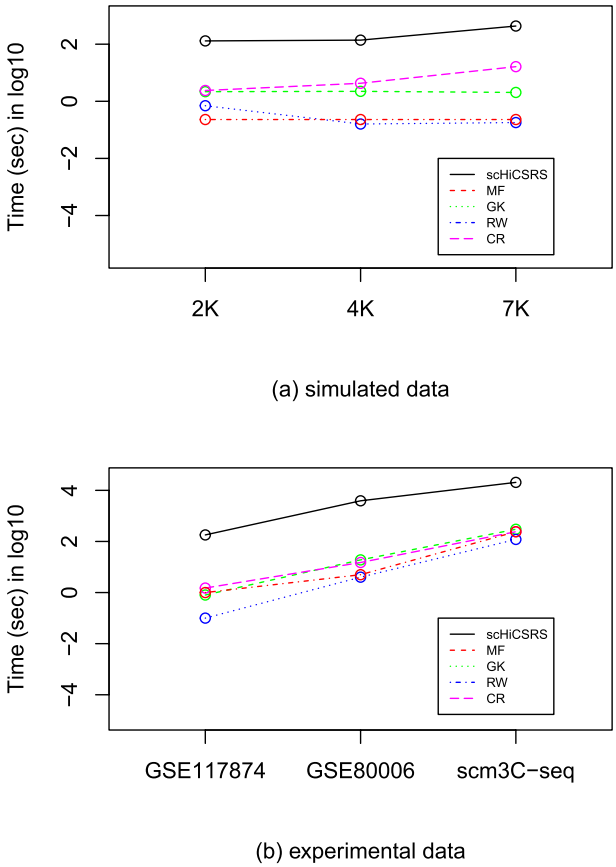

Fig. 5 Computation time of five methods for experimental and simulated data: **a** simulated data with 100 cells and three different sequencing depths: 2K, 4K, and 7K; **b** three sets of experimental data: GSE117874, GSE80006, and scm3C-seq

observed ones almost exactly. This provides the highest confidence in the imputed dropout values. For GSE80006, although there is an obvious diagonal line going through the pint cloud, the range of the imputed values is more and more inflated as the observed values get larger and larger; that is, the diagonal line going through the point cloud is not at 45 degree. This pattern is also seen for the scm3C-seq data, where the issue is further exacerbated by the scattering of the point cloud no longer tightly around a line, but rather around a band. To potentially remedy this problem, we will explore other normalization methods, including first applying BandNorm [37] for biophysical property confirmation before sequencing depth normalization to the median.

Another issue is related to computation, in particular, regarding the large memory requirement of scHiCSRS. As the dimension of the scHi-C contact matrix increases, the memory space it requires increases exponentially, making it difficult to run on a local computer. Further, scHiCSRS can be more computationally intensive compared to the other methods, especially when the number of cells analyzed together is large, as in the case of the L4 and L5 prefrontal cortex data (Fig. 5). This is not surprising given that, for scHiCSRS, all cells are analyzed simultaneously to borrow information from one another to increase statistical power and imputation accuracy, whereas the other methods analyze each cell separately. However, we believe that the fuller use of available information

and thus the much better performance of scHiCSRS justifies its computational cost, especially since it is still practically feasible. Nevertheless, effort will continue to be made to further improve the computational efficiency. One potential solution is to break a big dataset (say, with upward of a thousand of single cells) into smaller, manageable, subsets (say, in the lower hundreds) in the imputation steps. Then, all imputed values can then be analyzed jointly in the Gaussian mixture step for structural zeros inference. This could lead to the scalability of scHiCSRS although at a cost of some information. Another possible improvement is to use additional available information to enhance the contact matrices. In the current scHiCSRS model, we only borrowed information from the neighborhood and similar cells, although we could extend the formulation to incorporate information from bulk data. One possibility is to assume that the imputed count is a weighted average of neighborhood regions, similar cells, and bulk data.

Conclusion

This paper proposes scHiCSRS, which provides a valuable tool for identifying structural zeros and imputing dropouts. The resulted data are improved for downstream analyses, especially for understanding cell-to-cell variation through subtype clustering. Given that the observed zeros are a mixture of zeros that reflect true underlying biological mechanisms and those that are simply due to insufficient sequencing depth, the ability to tears out these two pieces is critically important for downstream analyses as we have demonstrated. If all zeros were treated as dropouts and imputed as small observed values, then significant differences in cells that are attributed to varying regulation mechanisms would be lost and their distinction would be much more difficult to ascertain. By calling structural zeros with high sensitivity, we greatly restore the data to better reflect the underlying true biological mechanism. In summary, since single cell Hi-C data are important for studying cell variability in gene regulation, and since it is commonly acknowledged that single cell Hi-C data need to be improved for better downstream analysis, we believe this work is timely and addresses a problem of scientific importance.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06147-8.

Additional file

Supplementary file 1

Acknowledgements
The authors would like to thank two anonymous reviewers for their insightful and constructive comments, which, we believe, has led to an improved manuscript.

Author contributions
QX developed the methodology, wrote the software, analyzed the data, and participated in writing the article. MW participated in data analyses, reviewed the results, and provided critical comments. SL conceptualized the ideas, supervised the methodological development and data analyses, and wrote the manuscript.

Funding
This research was supported in part by a grant from the National Institute of Health R01GM114142.

Availability of data and materials
The scHiCSRS R package, together with the processed experimental and simulated data used in this study, are available on Github at https://github.com/osu-stat-gen/scHiCSRS.git. For the experimental data, links to the websites and

accession numbers are: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117874, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80006, and GitHub - dixonlab/scm3C-seq.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no Conflict of interest.

Received: 8 July 2024 Accepted: 23 April 2025
Published online: 21 May 2025

References

1. Chen C, et al. scrmd: imputation for single cell RNA-seq data via robust matrix decomposition. bioRxiv. 2018;36:459404.
2. Claeskens G, et al. Model selection and model averaging. Cambridge: Cambridge Books; 2008.
3. Darrow EM, et al. Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. Proc Natl Acad Sci. 2016;113(31):E4504–12.
4. Dekker J. Gene regulation in the third dimension. Science. 2008;319(5871):1793–4.
5. Flyamer IM, et al. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature. 2017;544(7648):110–4.
6. Fraser J, et al. An overview of genome organization and how we got there: from fish to Hi-C. Microbiol Mol Biol Rev. 2015;79(3):347–72.
7. Han C, et al. Are dropout imputation methods for scRNA-seq effective for scHi-C data? Brief Bioinform. 2020;22:bbaa289.
8. Hong H, et al. DeepHiC: a generative adversarial network for enhancing Hi-C data resolution. PLoS Comput Biol. 2020;16(2):e1007287.
9. Hu Y, et al. Wedge: imputation of gene expression values from single-cell RNA-seq datasets using biased matrix decomposition. bioRxiv. 2020;22:864488.
10. Huang M, et al. Gene expression recovery for single cell RNA sequencing. bioRxiv. 2017. https://doi.org/10.1101/138677.
11. Jin K, et al. scTSSR: gene expression recovery for single-cell RNA sequencing using two-side sparse self-representation. Bioinformatics. 2020;36(10):3131–8.
12. Kim S, et al. The dynamic three-dimensional organization of the diploid yeast genome. Elife. 2017;6:e23623.
13. Lee D-S, et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. Nat Methods. 2019;16(10):999–1006.
14. Li WV, Li JJ. An accurate and robust imputation method scimpute for single-cell RNA-seq data. Nat Commun. 2018;9(1):1–9.
15. Lieberman-Aiden E, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.
16. Lyu H, et al. scHiCPTR: unsupervised pseudotime inference through dual graph refinement for single-cell Hi-C data. Bioinformatics. 2022;38(23):5151–9.
17. Mongia A, et al. Mcimpute: matrix completion based imputation for single cell RNA-seq data. Front Genet. 2019;10:9.
18. Nagano T, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59–64.
19. Nagano T, et al. Single-cell Hi-C for genome-wide detection of chromatin interactions that occur simultaneously in a single cell. Nat Protoc. 2015;10(12):1986.
20. Park J, Lin S. Evaluation and comparison of methods for recapitulation of 3D spatial chromatin structures. Brief Bioinform. 2019;20(4):1205–14.
21. Peng T, et al. Scrabble: single-cell RNA-seq imputation constrained by bulk RNA-seq data. Genome Biol. 2019;20(1):88.
22. Ramani V, et al. Sci-Hi-C: a single-cell Hi-C method for mapping 3D genome organization in large number of single cells. Methods. 2019;170:61–8.
23. Rao J, et al. Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. iScience. 2021;24(5):102393.
24. Rao SS, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
25. Rosenthal M, et al. Bayesian estimation of three-dimensional chromosomal structure from single-cell Hi-C data. J Comput Biol. 2019;26:1191.
26. Tan L, et al. Three-dimensional genome structures of single diploid human cells. Science. 2018;361(6405):924–8.
27. Ursu O, et al. Genomedisco: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics. 2018;34(16):2701–7.

28. van Dijk D, et al. Magic: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv. 2017. https://doi.org/10.1016/j.cell.2018.05.061.
29. Vershynin R. Introduction to the non-asymptotic analysis of random matrices. 2010*arXiv preprint* arXiv: 1011.3027.
30. Xiao G, et al. Modeling three-dimensional chromosome structures using gene expression data. J Am Stat Assoc. 2011;106(493):61–72.
31. Yang T, et al. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res. 2017;27(11):1939–49.
32. Zand M, Ruan J. Network-based single-cell RNA-seq data imputation enhances cell type identification. Genes. 2020;11(4):377.
33. Zhang Y, et al. Enhancing Hi-C data resolution with deep convolutional neural network HiCplus. Nat Commun. 2018;9(1):750.
34. Zhang Z. et al. Inference of spatial organizations of chromosomes using semi-definite embedding approach and hi-c data. In: annual international conference on research in computational molecular biology, Springer. 2013: 317–332
35. Zhao Y, et al. Link prediction for partially observed networks. J Comput Graph Stat. 2017;26(3):725–33.
36. Zhen C, et al. A novel framework for single-cell Hi-C clustering based on graph-convolution-based imputation and two-phase-based feature extraction. bioRxiv. 2021. https://doi.org/10.1101/2021.04.30.442215.
37. Zheng Y, et al. Normalization and de-noising of single-cell Hi-C data with BandNorm and SCVI-3D. Genome Biol. 2022;23(1):222.
38. Zhou J, et al. Robust single-cell Hi-C clustering by convolution-and random-walk-based imputation. Proc Natl Acad Sci. 2019;116:201901423.
39. Zhou X, et al. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning-based neural network. GigaScience. 2020;9(7):giaa076.
40. Zhu H, Wang Z. SCL: a lattice-based approach to infer 3D chromosome structures from single-cell Hi-C data. Bioinformatics. 2019;35(20):3981–8.

Publisher's Note